

INTRO TO DATA SCIENCE

LESSON 1: DATA EXPLORATION

INTRO TO DATA SCIENCE

WELCOME!

Instructor: Ed Podojil

TAs: Patrick McNamara, Michael Williams

Classes:

Tuesday, 6:30-9:30, 33 Union Square

Thursday, 6:30-9:30, GA East, Studio 1

Outside of classwork: Roughly 3 hours a week.

Office Hours: To Be Determined (Very Soon!)

Contact:

Patrick: patrick.mcnamara86@gmail.com

Michael: mvwilli@gmail.com

I. GOALS OF THE COURSE

II. WHAT IS DATA SCIENCE?

III. THE DATA MINING WORKFLOW

LAB:

IV. PYTHON SETUP

V. WORKING IN UNIX AND PYTHON

INTRO TO DATA SCIENCE

I. GOALS OF THE COURSE

Insight in how to understand and read data

Gain techniques for manipulating data

Learn machine learning algorithms commonly used in practice

Insight in how to understand and read data

Gain techniques for manipulating data

Learn machine learning algorithms commonly used in practice

Lots and lots of practice!

Generally lecture notes will follow these guidelines:

Class goals will almost always be labeled in a bold teal

`Code examples are monospaced and inverted`

Each section of class will end with review questions

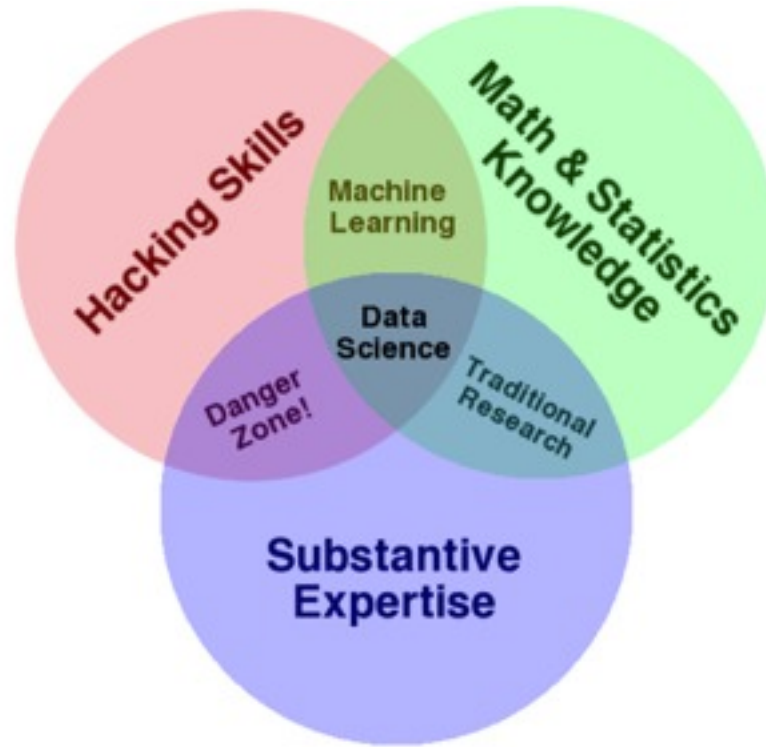
INTRO TO DATA SCIENCE

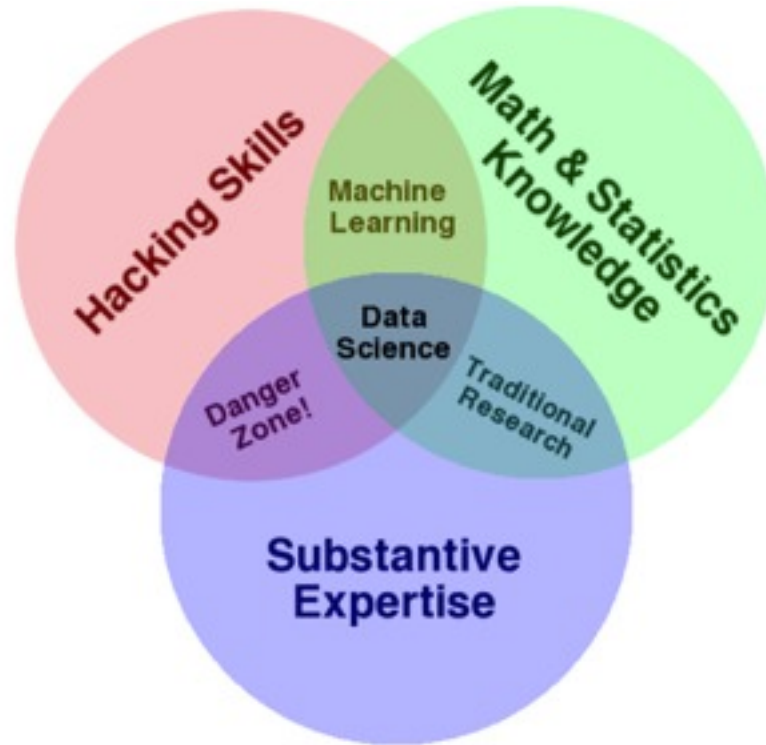
I. WHAT IS DATA SCIENCE?

A set of tools and techniques used to extract useful information from data.

A set of tools and techniques used to extract useful information from data.

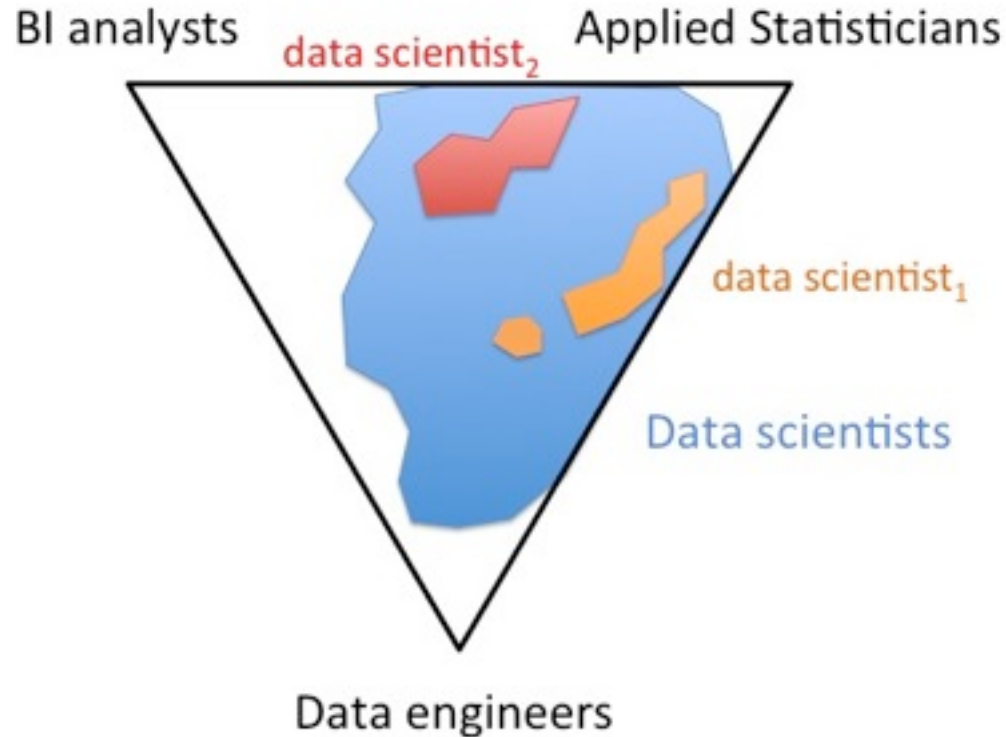
An interdisciplinary, problem-oriented subject.





ONE MORE THING!

Communication skills



A set of tools and techniques used to extract useful information from data.

An interdisciplinary, problem-solving oriented subject.

The application of scientific techniques to practical problems.

A set of tools and techniques used to extract useful information from data.

An interdisciplinary, problem-solving oriented subject.

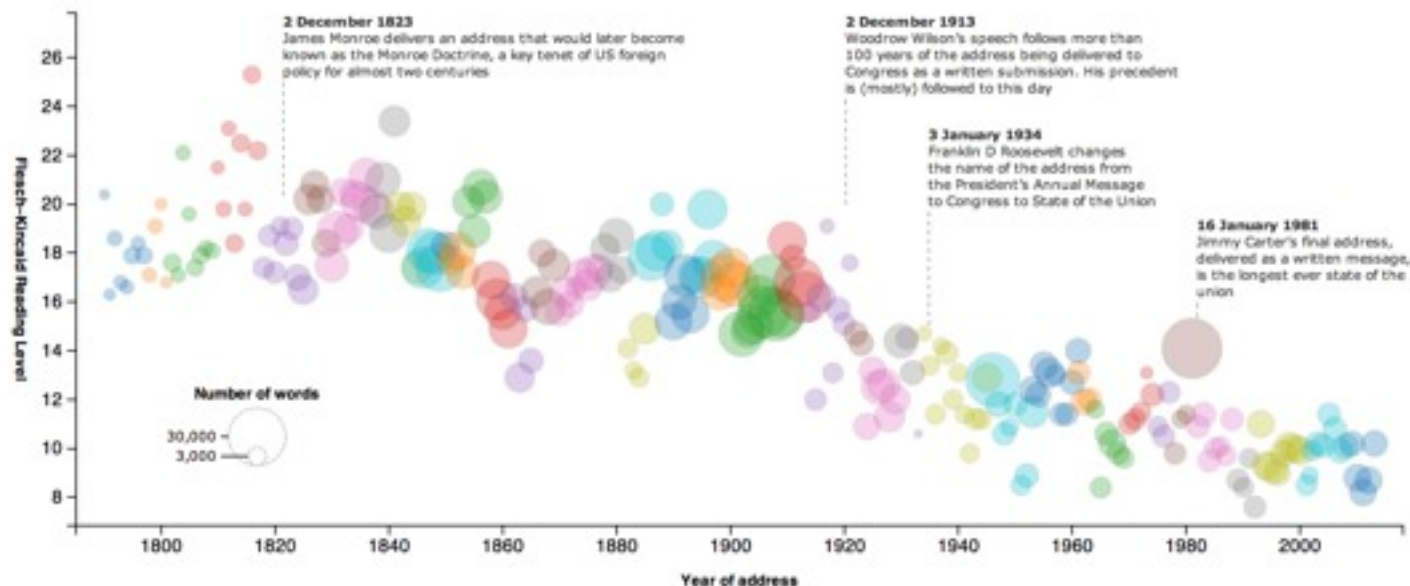
The application of scientific techniques to practical problems.

A rapidly growing field.



The state of our union is ... dumber: How the linguistic standard of the presidential address has declined

Using the [Flesch-Kincaid readability test](#) the Guardian has tracked the reading level of every state of the union



Music + Data:
<http://bit.ly/echonest>

- Stack Overflow tag recommendation and response time prediction
- Locating ethnic food in ethnic neighborhoods
- Building optimal fantasy football teams
- Recommending new musical artists
- Identifying key areas to get a taxi in NYC
- Finding the right job for you



Michael E. Driscoll

@medriscoll



Following

Data scientists: better statisticians than most programmers & better programmers than most statisticians bit.ly/NHmRqu
[@peteskomoroch](#)



Reply



Retweet



Favorite



More



Pocket

- Statistical and machine learning knowledge
- Computer Science experience (Applied Math)
- Academic curiosity
- Product sense
- Storytelling
- Cleverness

REVIEW

1. What are the leading qualities that make up a data scientist?
2. Name an example of a company that uses data science to help improve their product.

REVIEW

1. Creativity, a statistics and engineering background, wit
2. Amazon: Recommendations to get users to continue making purchases

II. THE DATA SCIENCE WORKFLOW

Dataists

1. Obtain
2. Scrub
3. Explore
4. Model
5. Interpret

Jeff Hammerbacher: Chief Scientist, Cloudera

1. Identify problem
2. Instrument data sources
3. Collect data
4. Prepare data (integrate, transform, clean, impute, filter, aggregate)
5. Build model
6. Evaluate model
7. Communicate results

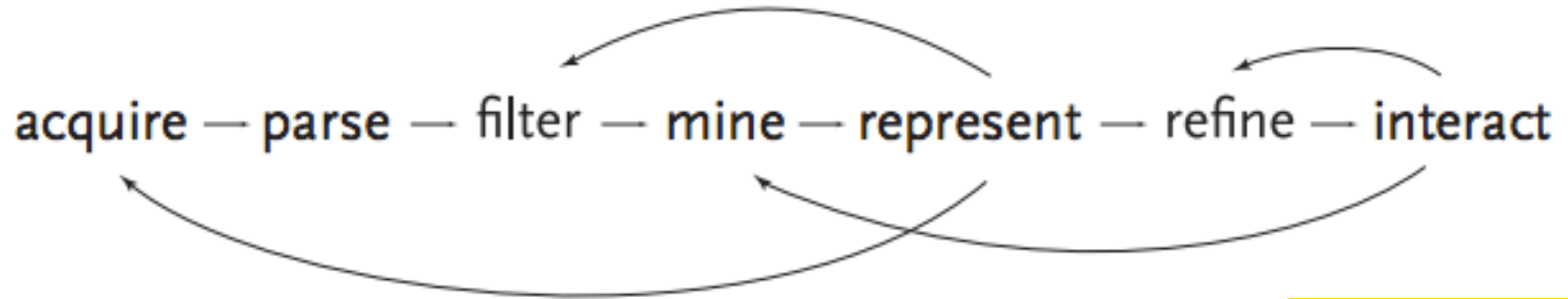
Ted Johnson: AT&T Research

1. Assemble an accurate and relevant data set
2. Choose the appropriate algorithm

Ben Fry: Principal, Fathom

1. Acquire
2. Parse
3. Filter
4. Mine
5. Represent
6. Refine
7. Interact





NOTE

This diagram illustrates the iterative nature of problem solving

PROBLEM: WHAT ARE THE LEADING INDICATORS THAT A USER WILL MAKE A NEW PURCHASE?

PROBLEM: WHAT ARE THE LEADING INDICATORS THAT A USER WILL MAKE A NEW PURCHASE?

1. Collect data around user retention, user actions within the product, potentially find data outside of company

PROBLEM: WHAT ARE THE LEADING INDICATORS THAT A USER WILL MAKE A NEW PURCHASE?

1. Collect data around user retention, user actions within the product, potentially find data outside of company
2. Extract aggregated values from raw data
 1. How many times did a user share through Facebook within a week? A month?
 2. How often did they open up our emails?

PROBLEM: WHAT ARE THE LEADING INDICATORS THAT A USER WILL MAKE A NEW PURCHASE?

1. Collect data around user retention, user actions within the product, potentially find data outside of company
2. Extract aggregated values from raw data
 1. How many times did a user share through Facebook within a week? A month?
 2. How often did they open up our emails?
3. Examine data to find common distributions and correlations

PROBLEM: WHAT ARE THE LEADING INDICATORS THAT A USER WILL MAKE A NEW PURCHASE?

1. Collect data around user retention, user actions within the product, potentially find data outside of company
2. Extract aggregated values from raw data
 1. How many times did a user share through Facebook within a week? A month?
 2. How often did they open up our emails?
3. Examine data to find common distributions and correlations
4. Extract new meaning to predict if a user would purchase again or not

PROBLEM: WHAT ARE THE LEADING INDICATORS THAT A USER WILL MAKE A NEW PURCHASE?

1. Collect data around user retention, user actions within the product, potentially find data outside of company
2. Extract aggregated values from raw data
 1. How many times did a user share through Facebook within a week? A month?
 2. How often did they open up our emails?
3. Examine data to find common distributions and correlations
4. Extract new meaning to predict if user would purchase again
5. Share results (and probably also go back to the drawing board)

REVIEW

1. What seem to be the most common practical steps in the data science workflow?
2. Is the workflow straightforward? Why or why not?

REVIEW

1. Collect data, explore data, create a model, share the results
2. Usually, no! There will always be a need to collect more data and improve the original model.

IV. PYTHON SETUP

High variety of languages used in practice

Statistics: “Python,” R, Matlab, Julia, Fortran, STATA

Scripting: Python, Ruby, Scala, Java

Data Querying: SQL, Hive, Pig

Python is an open source project which is maintained by a large and very active community.

Python is an open source project which is maintained by a large and very active community.

It was originally created by Guido Van Rossum in the 1990s, who currently holds the title of Benevolent Dictator For Life (BDFL).

The presence of a BDFL means that Python has a unified design philosophy.

This design philosophy emphasizes readability and ease of use, and is codified in PEP8 (the Python style guide) and PEP20 (the Zen of Python).

NOTE

PEPs (or Python Enhancement Proposals) are the public design specs that the language follows.

Python in nature is not a statistical language, though used in a variety of ways: web applications, server maintenance, reading and writing text files:

web development
systems admin
(etc)

<https://www.djangoproject.com/>

<http://docs.fabfile.org/en/1.6/>

<https://github.com/languages/Python>

Python evolved alongside Bioinformatics and Data Analysis, introducing stats packages (numpy, scipy) and machine learning packages (scikit-learn, NLTK)

ADVANTAGES

- VERY FAST COMPARED TO R
- USEFUL ACROSS PLATFORMS
- EASY TO INTEGRATE
- COMMON OOP TECHNIQUES
- GREAT DOCUMENTATION SUPPORT

DISADVANTAGES

- NO GREAT VISUALIZATION PACKAGES (YET!)
- NATURAL DISPLAY IS LESS READABLE
- LESS NEWBIE FRIENDLY
- LACK OF PARALLEL PROCESSING

III. PYTHON DATA STRUCTURES

The most basic data structure is the None type. This is the equivalent of NULL in other languages.

There are four numeric types: **int**, **float**, **bool**, **complex**.

```
>>> type(1)
<type 'int'>
>>> type(2.5)
<type 'float'>
>>> type(True)
<type 'bool'>
>>> type(2+3j)
<type 'complex'>
```

The next basic data type is the Python list.

A list is an ordered collection of elements, and these elements can be of arbitrary type. Lists are mutable, meaning they can be changed in-place.

```
>>> k = [1, 'b', True]
>>> k[2]
True
>>> k[1] = 'a'
>>> k
[1, 'a', True]
```

Likewise, **tuples are immutable arrays of arbitrary elements.**

```
>>> x = (1, 'a', 2.5)
>>> x
(1, 'a', 2.5)
>>> x[0]
1
>>> x[0] = 'b'
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: 'tuple' object does not support item assignment
```

Tuples are frequently used behind the scenes in a special type of variable assignment called tuple packing/unpacking.

The string type in Python represents an immutable ordered array of characters (note there is no char type).

Strings support slicing and indexing operations like arrays, and have many other string-specific functions as well.

String processing is one area where Python excels.

Associative arrays (or hash tables) are implemented in Python as the dictionary type.

```
>>> this_class = {'subject': 'data science', 'instructor': 'jason', 'time': 1800, 'is_cool': True}
>>> this_class['subject']
'data science'
>>> this_class['is_cool']
True
```

Dictionaries are unordered collections of key-value pairs, and dictionary keys must be immutable.

Another basic Python data type is the set. Sets are unordered mutable collections of distinct elements.

```
>>> y = set([1,1,2,3,5,8])  
>>> y  
set([8, 1, 2, 3, 5])
```

These are particularly useful for checking membership of an element and for ensuring element uniqueness.

INTRO TO DATA SCIENCE

LAB: UNIX AND PYTHON

IN CLASS WORK

1. Change our python script to also return minimum, maximum, and average age, and click through rate (clicks/impressions)
 2. Upload your working script to Schoology when you're done
- EXTRA CREDIT:** Update the script to write a new file instead of using standard out.

INTRO TO DATA SCIENCE

DISCUSSION

FOR NEXT TIME:

1. Find a data science link you want to share with the class.
This can be a blog, paper worth reading, or article you found.
2. Post this link on Schoology and share it with the class!

NEXT CLASS SUBJECT: LINEAR ALGEBRA